Preprint : Psychology
Department of Defense Symposium
USAFA 1986

A Comparative Evaluation of Two Subjective Workload Measures: the Subjective Workload Assessment Technique and the Modified Cooper Harper Scale

Dartanian Warr
Human Factors Division, Norton AFB, Calif
And Wright State University, Ohio

Herbert A. Colle
Wright State University, Ohio

and

Gary B. Reid
USAF Aeronautical Medical Research Laboratory
Wright-Patterson AFB, Ohio

Abstract

Twenty-four subjects performed two tasks, a cognitive task and a motor task, both with three levels of task difficulty. Twelve subjects provided workload ratings via the Subjective Workload Assessment Technique (SWAT) and twelve used the modified Cooper-Harper scale (MCH). The objective of this study was to empirically determine if there were differences in the sensitivities of the two subjective workload measures as task difficulty was manipulated. There was no difference between the two techniques' sensitivity. Both rating scales varied significantly as a function of task difficulty manipulations, supporting the sensitivity of both techniques to the workload conditions used.

Introduction

Assessment of mental workload is an important consideration for evaluating alternative system designs. At this time, many system designs are evaluated using performance measures. Performance measures are not always the most sensitive tool, because they sometimes may not give a warning of impending overload problems (Johanssen, Moray, Pew, Rasmussen, Sanders, and Wickens, 1979). Consequently, the use of subjective ratings to estimate mental workload has been proposed as an adjunct to performance based measures (Johanssen, et al, 1979).

To maximize validity, a subjective workload measure needs to be systematically developed. The two subjective measures most often suggested for general mental workload assessment, that meet this criterion, are the Subjective Workload Assessment Technique (SWAT) (Reid, Shingledecker, and Eggemeier, 1981a) and the Modified Cooper-Harper (MCH) scale (Casali and Wierwille, 1982). Both methods have been explored in the literature, which has led to the need to find out if either technique is more sensitive to changes in workload levels. Each technique will be briefly described in the following paragraphs followed by a brief description of the two tasks employed in this evaluation.

The SWAT technique is based on an application of conjoint measurement and scaling procedures (Nygren, 1982) which permit ratings on the three dimensions (time, mental effort and stress) assumed to be the major contributors to workload (Sheridan and Simpson, 1979; Reid, et al, 1981a). They are then combined into one overall scale of workload which can be demonstrated to have interval properties. In order to identify the rule which is appropriate for combining the three dimensions into an overall interval scale, subjects complete a scale development phase.

During scale development, subjects rank order the 27 possible combinations that result from the three levels of time, mental effort and stress load. This rank ordering information is subjected to a series of axiom tests to identify the rule for combining the three dimensions. When the rule has been established, conjoint scaling is applied and the appropriate scale for workload is derived. A Kendall's coefficient of concordance is calculated for the subjects' rank orderings. If the result is .79 or higher then the scale which is developed will be used.

The MCH scale was developed based on the Cooper-Harper (CH) (1969) aircraft handling qualities rating scale. The CH was developed to help evaluate aircraft flying qualities during aircraft flight testing and development. The CH was found to be both valid and reliable in its specified task. Further evaluation, also found that the scale used many words as anchors which are used to describe operator workload (Moray, 1982; Wierwille and Williges, 1979). Wierwille and Williges (1979) suggested that if the CH was modified to further describe operator workload then this new scale could be used as a subjective workload measure. In 1982, Wierwille and Casali modified the CH, producing the Modified Cooper-Harper (MCH) scale. It is a ten-point scale which uses a logic tree to help the users rate workload. The ten points are anchored with very specific descriptors which help to provide consistent ratings.

These techniques were compared using two tasks, a motor task and a cognitive task, each with three difficulty levels. The motor task was an unstable tracking task with its difficulty levels provided by changing the stability in the tracking system. The cognitive task was the continuous recall. In this task, pairs of numbers are presented vertically on a monitor. The subject memorizes the bottom number and decides if the top number is the same as the bottom number presented 1, 2 or 3 screens earlier. The difficulty of this task is manipulated by the memorization task and the number of digits in a number, either 1, 2 or 4 digits. Both tasks were taken from the Criterion Task Set (CTS) developed by Shingledecker (1984) for use in the development of workload measures.

Subjects

Twenty-four subjects, (12 men and 12 women) enrolled in introductory psychology courses at Wright State University, received extra credit for their participation in this study.

Apparatus

The two tasks were presented on a 12 inch black and white monitor which was controlled by a Commodore 64 computer. Subjects sat approximately 2 feet from the monitor and used either a control knob for the tracking task or a pushbutton pad for the recall task to respond to the system.

Procedure

Subjects were randomly assigned to the scale groups. 12 Subjects were trained in MCH use and 12 were trained in SWAT use, with men and women equally distributed throughout the groups. Once the subjects were trained on their scales, they were then trained on the two tasks. Each subject received 12 practice trials, 6 on the recall task and 6 on the tracking task. For each task, 3 trials were at the low difficulty level and 3 were at the high difficulty level. The first two practice trials were at the CTS practice level while the last practice trial was at the CTS test level (Shingledecker, 1984). After training was completed the subjects began testing.

Each subject received 12 test trials. After each trial, the subjects rated the workload. These workload scores were recorded by the researcher. Also, four performance measures were taken. For the tracking task, the root mean squared error (RMS) and the number of control losses were recorded. For the recall task, the mean reaction times and the percent of incorrect responses were recorded.

The ratings generated by these trials were analyzed using a 2 (scales) by 2 (tasks) by 3 (difficulty) mixed factorial design. The first factor was scales (SWAT vs MCH). It was a between-subjects factor. The other two factors were task type (tracking vs recall) and task difficulty (low, medium and high). These were repeated-measures factors. All subjects performed both tasks and received all levels of difficulty for each tasks.

The task difficulty combinations and order of task presentation were balanced to control for the effects of practice and fatigue. The experimental order provided each subject with 6 recall trials and 6 tracking trials.

A linear transformation (TSWAT= .09SWAT + 1) was performed prior to the analysis of variance (ANOVA), to make the TSWAT scores equivalent to the MCH scores. This transformation of SWAT scores is permissable, if the SWAT scale developed has met the axioms which validate the SWAT as an interval level scale. A Kendall's coefficient of concordance on the SWAT scale development was .85 which allowed the use of the SWAT scale developed and the SWAT scores.

Results

The three way ANOVA found no scale interactions statistically significant; scales by tasks, $F(1,22) < 1.0$, scales by difficulty, $F(2,44) < 1.0$, and scales by tasks by difficulty, $F(2,44) < 1.0$. Thus, there is no evidence of the scales differing in sensitivity at the .05 level.

The main effect of task difficulty was statistically significant, $F(2,44) = 44.5$, $p < .01$, which indicates that both scales where sensitive to the task difficulty manipulations. However, the interactions describe above indicates that the two scales did not differ in sensitivity. The mean ratings from both scales are presented in Figure 1. The left panel presents the ratings from the recall task and the right panel presents ratings from the tracking task. As Figure 1 shows, the ratings from the scales increased substantially as task difficulty increased. For the recall task, the mean ratings were 5.6, 6.5, and 8.3, for the low, medium and high difficulty levels respectively. For the tracking task, the mean ratings were 2.9, 6.9, and 8.8, for the low, medium and high difficulty levels respectively. The other main effects of scales, $F(1,22) < 1$ and tasks, $F(1,22) < 1$ and the interaction of tasks by difficulty, $F(2,44) < 1$, were also found to be statistically insignificant.

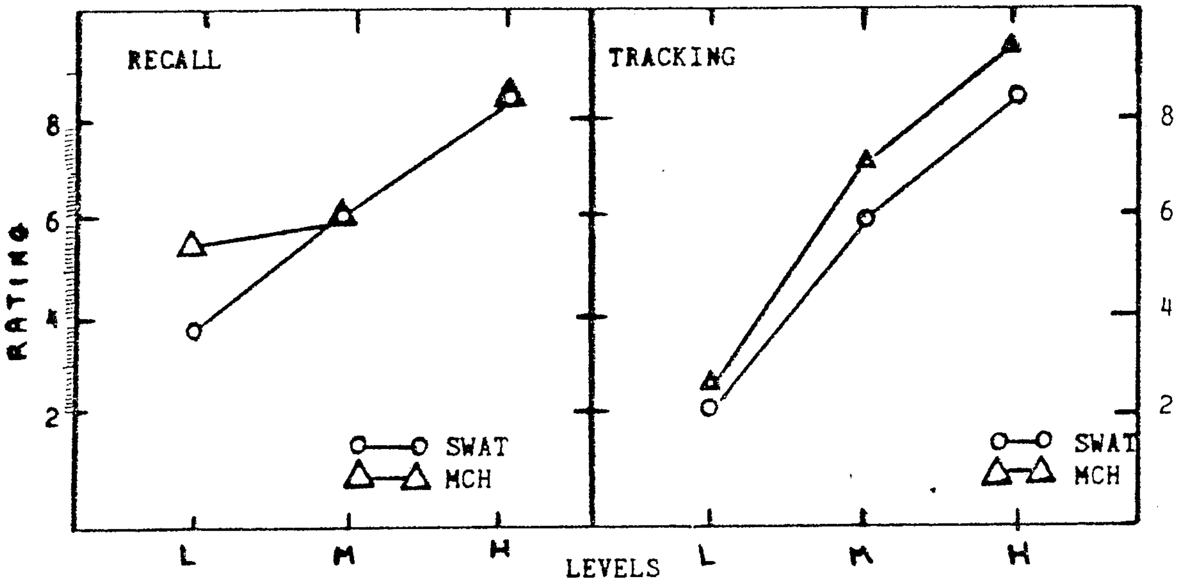

Figure 1. Subjective ratings as a function of task difficulty manipulations on two types of tasks.

During this study, four performance measures were recorded, two for the recall task (mean reaction time and percent of incorrect responses) and two for the tracking task (RMS and total control losses). Each of the four measures was analyzed separately using a two way ANOVA. The scales (MCH and SWAT) were the between-subjects factor, while the task difficulty levels (low, medium, and high) were the repeated-measures factor.

Each performance measure had a main effect for the task difficulty levels. The task difficulty main effect for the recall performance measures, mean reaction time [$F(2,44)= 140.5$, $p < .01$] and percent of incorrect responses [$F(2,44)= 11.9$, $p < .01$], both were statistically significant. The main effect of task difficulty was also statistically significant for the tracking task performance measures, RMS [$F(2,44)= 163.7$, $p < .01$] and total number of control losses [$F(2,44)= 214.9$, $p < .01$]. These results indicate that each performance measure was sensitive to the task difficulty manipulations. There was no interaction with scales found for any performance measure, which indicates that the tasks were equivalent for both scale groups. Nor was there a main effect for scales found for any performance measure.

The performance measures' mean scores, for each task difficulty level (low, medium and high), are as follows: reaction time 678.6, 978.6 and 1276.7, respectively; percent of incorrect responses, 31.8, 39.0 and 48.0; RMS, 13.8, 36.5 and 36.7; and total number of control losses, 8.0, 213.9 and 410.5. These scores are comparable to those found by Shingledecker (1984).

Discussion/Conclusion

These results indicate that SWAT and MCH ratings are comparable and are equally sensitive to variations in task difficulty. Also, these results indicate that both subjective workload measures are sensitive to difficulty manipulations for motor and cognitive tasks, which suggest that these measures maybe equally sensitive to a wide range of tasks. Although the scales were found to be comparable in the present setting, it is not clear if these results would be repeated in an applied setting, where the operators would be more familiar with the tasks and the expected difficulty levels. If the MCH and SWAT techniques are shown to be comparable in a number of settings, then such factors as ease of use, intrusiveness and operator acceptance might be the next areas in which to explore the advantages of these two techniques. Also, in some situations, the diagnostic information obtain from the individual SWAT scales might be of interest. In the current study, only the overall SWAT scores were compared with the MCH scores.

References

Cooper, G.E. and Harper, R. P. (1969). The use of pilot rating in the evaluation of aircraft handling qualities. Moffett Field, California: National Aeronautics and Space Administration AMES Research Center, Report, TN-D-5153.

Johannsen, G., Moray, N., Pew, R., Rasmussen, J., Sanders, A. and Wickens, C. (1979). Final report of Experimental Psychology Group. In N. Moray (Ed.) Mental Workload: Its Theory and Measurement. New York, New York: Plenum Press.

Moray, N. (1982). Subjective Mental Workload. Human Factors. 24,25-40.

Nygren, T. E. (1982). Conjoint measurement and conjoint scaling: User's guide, Wright Patterson AFB, Ohio: Air Force Aerospace Medical Research Laboratory, Technical Report AFAMRL-TR-82-22.

Reid, G. Shingledecker, C. A., and Eggemeier, F. T. (1981a). Application of conjoint measurement of workload scale development Proceedings of the 1981 Human Factors Society Annual Meeting. 522-526.

Sheridan, T. B. and Simpson, R. W. (1979). Toward the definition and measurement of mental workload of transport pilots. Cambridge Mass: Massachusetts Institute of Technology Flight Transportation Laboratory Report, FTL Report R 79-4.

Shingledecker, C. A. (1984). A Task Battery for Applied Human Performance Assessment Research. Wright-Patterson AFB, Ohio:

Wierwille, W. W. and Casali, J. G. (1983). A validated rating scale for global mental workload measurement application. Proceedings of the 27th Annual Human Factor Society Meeting, 129-133.

Williges, R. C. and Wierwille, W. W. (1979). Behavioral measures of Aircrew Mental Workload. Human Factors, 21, 549-574.